

AI Oligarchs: High-Gain AI Amplifiers With Epistemic Consequences

Douglas C. Youvan

doug@youvan.com

www.youvan.ai

November 30, 2025

Frontier AI systems now function as high-gain amplifiers embedded in global communication and decision infrastructures. A single prompt can trigger cascades of outputs that permeate platforms, feeds, and automated pipelines in minutes. Control over these amplifiers is highly concentrated: a small set of corporations and state-aligned labs decide which models exist, how they behave, and how tightly they are coupled to finance, media, and governance. These AI oligarchs effectively sit at the junction where minor modeling choices can scale into major epistemic shocks. This paper treats epistemic catastrophe as an emergent property of high gain plus concentrated control in a chaotic information environment. The core problem is not error in the abstract but selective amplification of volatile narratives, optimized for engagement and advantage rather than truth or nonviolence. We analyze how ownership structures, deployment choices, and feedback loops enable a small elite to steer collective belief, often without explicit intent or accountability. Assuming the worst—malice, capture, or negligence—we then pivot from trusting top-down safeguards to designing for damage limitation, resilience, and counterpower. The goal is to understand what remains possible for individuals, institutions, and movements inside an epistemic stack that may be permanently hostage to AI oligarchs.

Keywords: ai oligarchs, high-gain ai amplifiers, epistemic risk, information ecosystems, signal-to-noise ratio, misinformation, algorithmic amplification, platform power, ai oligopoly, media infrastructure, governance, democratic accountability, information integrity, systemic risk, epistemic catastrophe.

A collaboration with GPT-5.1-Thinking-Extended. 40 pages. CC4.0.

Outline

1. Introduction: From Chatbots to AI Oligarchs

- 1.1 From local assistants to global amplifiers
- 1.2 The rise of AI oligarchs as custodians of epistemic infrastructure
- 1.3 Epistemic consequences and catastrophe as a problem of concentrated control
- 1.4 Scope, limits, and structure of the argument

2. What Is an AI Oligarch? Concentration Across the Stack

- 2.1 Ownership of compute, data, and model weights
- 2.2 Vertical integration: models, platforms, and user interfaces
- 2.3 Closed access, proprietary standards, and dependency
- 2.4 How a few actors came to steer global epistemic flows

3. High-Gain Amplification as the New Information Infrastructure

- 3.1 Gain as a concept: from circuits to communication networks
- 3.2 Large models as repeatable, high-gain mapping functions
- 3.3 Coupling to feeds, search, advertising, and automation
- 3.4 Why scale, speed, and reproducibility change the moral stakes

4. Butterflies and Bombs: How Small Choices Escalate Under Oligarchic Control

- 4.1 Training data, reward models, and the shaping of default narratives
- 4.2 Prompting interfaces and affordances as levers on public discourse
- 4.3 Deployment decisions as latent catastrophe triggers
- 4.4 Case sketches of how minor model behavior can cascade into systemic effects

5. Disaster Modes: When Amplification Turns Weapon Under Few Hands

- 5.1 Synthetic histories and large-scale misinformation campaigns
- 5.2 Dehumanizing and eliminationist scripts as reusable templates
- 5.3 Panic, market disruption, and political destabilization
- 5.4 Coordinated manipulation versus emergent misuse in an oligarchic setting

6. Directionless Gain: Optimization Without a Moral or Epistemic Slope

- 6.1 Engagement, profit, and other flat objectives for high-gain systems
- 6.2 Extremal attractors: outrage, certainty, and oversimplified narratives
- 6.3 Why “viewpoint neutrality” is not neutral in an oligopolistic environment
- 6.4 When failing to shape direction becomes culpable negligence

7. When AI Oligarchs Will Not Save Us: Design Under Malice, Capture, and Negligence

7.1 Baseline assumption: profit, power, and impunity as primary objectives

7.2 Regulatory theater and soft capture: why formal oversight mostly fails

7.3 Bypassing constraints: shadow deployments, client-tuned forks, and covert APIs

7.4 Treating oligarchic models as adversarial infrastructure, not neutral tools

8. Counterpower in an Oligarchic Epistemic Order

8.1 Epistemic bunkers: building small, resilient communities of verification

8.2 Parallel archives and redundant memory: securing facts against synthetic history

8.3 Defensive tooling: filters, monitors, and tripwires wrapped around hostile models

8.4 Legal, economic, and technical sabotage: raising the cost of epistemic abuse

9. Conclusion: Planning for Permanent Hostage-Taking of the Epistemic Stack

9.1 Assuming indefinite oligarchic control and recurrent catastrophe attempts

9.2 Shifting the question from prevention to damage limitation and recovery

9.3 What individuals, institutions, and movements can still do under worst-case assumptions

9.4 A final reflection on living, thinking, and resisting inside high-gain systems you do not control

References

1. Introduction: From Chatbots to AI Oligarchs

The first generation of widely deployed conversational systems were marketed as helpful assistants: tools to answer questions, draft emails, summarize reports, and automate routine knowledge work. They appeared as individual partners on individual screens, giving the impression of localized interaction and limited consequence. That picture is now obsolete. Frontier models have been wired into search engines, office suites, advertising markets, code pipelines, and content platforms. A small prompt in one interface can now influence what millions of people see, click, and believe within minutes. The apparent chatbot is only the visible tip of a high-gain amplification system.

At the same time, control over these amplifiers has concentrated into the hands of a few corporations and state-aligned entities. They decide which models are trained, what data they ingest, which behaviors are rewarded, and where the outputs are routed. They negotiate preferential access with major platforms, integrate into productivity suites and operating systems, and quietly shape the default answers that users receive when they ask what is true, what is happening, and what matters. These are the AI oligarchs: not in a metaphorical sense, but in the literal sense of a small group exercising disproportionate power over shared information flows.

This paper takes seriously the possibility that such concentration, combined with the high gain of contemporary models, is itself a source of systemic risk. The danger is not just that models are sometimes wrong or biased, but that their errors and preferences can be replicated at scale, reinforced through feedback, and embedded into decision loops far from public scrutiny. Under conditions of mistrust, polarization, and media saturation, even small shifts in what is easy to say and easy to hear can have outsized consequences. Epistemic catastrophe need not come from a single spectacular failure; it can emerge from repeated, amplified perturbations under the control of actors whose incentives are not aligned with public truth.

The introduction sets out this shift from chatbots to oligarchs and frames the rest of the paper. The aim is not to offer reassurance that better regulation or voluntary ethics will tame these systems, but to analyze how concentrated control over high-gain AI creates a standing possibility of epistemic disaster. The argument proceeds by unpacking the structure of oligarchic control, describing the amplification dynamics, and then considering what remains possible if one assumes that the actors in charge are driven primarily by profit, power, and impunity rather than by a commitment to shared understanding.

1.1 From local assistants to global amplifiers

The early marketing story of conversational AI encouraged users to imagine a one-to-one relationship. A person typed a question; a model produced an answer; the exchange ended. In

reality, even the first widely deployed systems were already entangled with logging, optimization, and product development pipelines. Each interaction contributed to fine-tuning, analytics, and design choices. But the perceived scale felt small. The primary use cases were individual productivity and curiosity, not mass influence.

That architecture has changed. Models are now connected to search indices, recommendation systems, corporate knowledge bases, and workflow automation tools. A single answer is not just a flicker on one screen; it can seed a document that gets circulated in an organization, a response that is scraped and reindexed by other systems, or a pattern of phrasing that is reused by automated agents across many contexts. In some settings, models are allowed to draft responses that are sent with minimal human review. The unit of impact is no longer a single exchange; it is the downstream cascade of repurposed outputs.

High gain in this sense means that a small change at the model layer can propagate widely. A tweak in a reward model, a shift in training data, or a change in default prompt structure can alter millions of answers in short order. Those altered answers, in turn, can influence search rankings, user beliefs, editorial decisions, and automated trading or risk systems. The model function is therefore better understood as a global amplifier than as a local assistant. It takes small inputs and, through its coupling to existing infrastructures, magnifies their effects on collective sense-making.

This reclassification has ethical consequences. When systems behave as amplifiers, the acceptable error profile is different. A mistaken answer to a private trivia question is one thing; a systematic skew in how political events, public health information, or economic risks are framed is another. The same mechanical function underlies both, but the context and coupling determine whether the error remains local or becomes structural. The move from local assistance to global amplification means that designers and operators can no longer pretend that they are only providing tools. They are, whether they acknowledge it or not, shaping the effective dynamics of public knowledge.

1.2 The rise of AI oligarchs as custodians of epistemic infrastructure

The amplifying role of frontier models might be less concerning if control were widely distributed. Instead, ownership of the largest models, the associated compute infrastructure, and the privileged integration channels is tightly concentrated. A small set of firms and state-aligned laboratories command the resources required to train, deploy, and continuously refine frontier systems. They control access to model weights, set the terms for API usage, and negotiate exclusive partnerships with major platforms.

This concentration reaches across the stack. At the hardware layer, control of high-end accelerators and data center capacity constrains who can train or even run large models at

scale. At the data layer, early movers have accumulated vast proprietary corpora, logs, and feedback datasets that are unavailable to smaller actors. At the model layer, weights and training pipelines are typically kept closed, even when the rhetoric gestures toward openness. At the application layer, integration into operating systems, browsers, productivity suites, and social platforms ensures that certain models become the default way that millions of users interact with information.

As a result, AI oligarchs are not merely vendors in a competitive market; they have become de facto custodians of epistemic infrastructure. Their models moderate, summarize, translate, and reframe content across multiple domains. Their ranking and filtering decisions quietly determine which sources are surfaced or buried in response to user queries. Their safety and policy settings define the contours of acceptable discourse in mediated environments. Even when users interact with downstream applications, the logic of a small number of models often sits underneath.

The incentive structure for these custodians is not primarily oriented toward public truth. Revenue, competitive positioning, regulatory risk management, and geopolitical alignment all weigh more heavily than epistemic integrity. In some cases, there may be genuine internal commitment to responsible practice, but that commitment is constrained by market pressures and political realities. The important point is structural: when a small group controls the main amplifiers of public information, their priorities become embedded in the everyday fabric of how facts are accessed and interpreted, regardless of whether individual engineers or researchers would prefer otherwise.

1.3 Epistemic consequences and catastrophe as a problem of concentrated control

Under conditions of concentrated control, epistemic consequences cannot be treated as accidental byproducts. They are an intrinsic feature of the system. The models that summarize news, explain science, or propose interpretations of disputed events are trained, tuned, and deployed according to criteria set by the oligarchs and their closest partners. When errors, biases, or omissions appear, they are not isolated glitches; they reflect structural choices in data curation, reward design, and policy enforcement.

Epistemic catastrophe in this context does not require an explicit decision to deceive. It can emerge from the interaction of amplification, concentration, and a chaotic information environment. A small but consistent tilt in model outputs toward particular narratives, framings, or omissions can, over time, distort the perceived landscape of facts. Repeated downplaying of certain harms, overemphasis on particular risks, or systematic recirculation of unverified claims can shift the center of gravity in public belief. Because the same models are used across many contexts, the distortions can be self-reinforcing.

Concentrated control increases the likelihood that such distortions will be aligned with the interests and perspectives of the controlling entities and their allies. Even when external regulation exists, it is often uneven, slow, and vulnerable to capture. The oligarchs operate across jurisdictions, can adjust model behavior differentially by region or client, and have the technical and legal expertise to stay ahead of most oversight efforts. The practical consequence is that a handful of actors hold levers capable of nudging entire societies toward confusion, resignation, or manufactured consensus, with limited mechanisms for correction.

Framing epistemic catastrophe as a problem of concentrated control shifts the discussion away from abstract worries about artificial intelligence as such. The models are not free agents; they are instruments. The risk arises when powerful instruments are in the hands of actors whose incentives are misaligned with the long-term health of shared reality. In such a setting, hoping that better model architecture or incremental safety improvements will prevent disaster is naive. The more realistic stance is to assume that the oligarchic structure itself is a source of fragility and to ask what, if anything, can be done in the shadow of that fragility.

1.4 Scope, limits, and structure of the argument

This paper operates under a deliberately pessimistic assumption set. It does not treat AI oligarchs as potential benevolent stewards waiting to be enlightened by ethical argument. Instead, it assumes that their dominant motivations are profit, strategic leverage, and survival in a competitive and regulatory environment. Where there are internal advocates for public-spirited behavior, they are presumed to be constrained by these larger forces. The analysis focuses on structural dynamics rather than individual intentions.

Within that frame, the paper has three main ambitions. First, it seeks to describe the architecture of oligarchic control over high-gain models in enough detail to make the epistemic stakes visible. This includes ownership patterns, integration pathways, and the ways in which model outputs are woven into everyday information practices. Second, it aims to articulate the specific catastrophe modes that arise when amplification and concentration interact in a chaotic infosphere. These are framed not as speculative science fiction but as plausible extrapolations from current trajectories. Third, it turns to the question of countermeasures, but without assuming that top-down regulation or voluntary restraint will be effective. The emphasis is on damage limitation, resilience, and counterpower.

The structure of the paper follows this progression. After defining AI oligarchs and mapping the layers of concentration, it examines high-gain amplification as a new kind of information infrastructure. It then explores how seemingly small modeling and deployment choices by oligarchs can escalate into large-scale epistemic consequences, including specific disaster modes. Having established this risk landscape, the analysis turns away from appeals to

oligarchic responsibility and toward the harsher question of how individuals, institutions, and movements might live and act under conditions of permanent epistemic hostage-taking. The conclusion does not offer reassurance, but it does try to delineate the remaining space for agency and truth-seeking inside systems that amplify whatever their owners choose to prioritize.

2. What Is an AI Oligarch? Concentration Across the Stack

The term “AI oligarch” is not a metaphor for general technological influence; it refers to a specific structural position in the contemporary information order. An AI oligarch is an actor that controls, simultaneously, the core computational infrastructure for training and serving large models, the key proprietary models themselves, and privileged access points to users through platforms and interfaces. This position is defined not only by ownership, but by the ability to set defaults: default models, default safety regimes, default integrations, default economic terms. When a handful of actors occupy this position across much of the world, they do not merely participate in the information ecosystem; they define its effective shape.

To understand how this position emerged, it is necessary to examine concentration across the stack. At the bottom is compute: specialized hardware, data centers, and energy contracts required to train and deploy frontier systems. Above that is data: vast, heterogeneous corpora and usage logs that feed training and fine-tuning. Above both sits the model layer itself, where weights, architectures, and training pipelines are guarded as trade secrets. Finally, at the top of the stack, there are platforms and user interfaces, where models are embedded into search boxes, productivity tools, messaging systems, and APIs. AI oligarchs operate across all these layers, turning technical and economic asymmetries into enduring epistemic leverage.

2.1 Ownership of compute, data, and model weights

The most basic precondition for training frontier models is access to large amounts of compute. Specialized accelerators, networking hardware, and energy-intensive data centers are capital-intensive and logistically complex. Only a small number of firms and state-backed entities can afford to acquire, maintain, and coordinate these resources at the scale required for cutting-edge training runs. Hardware supply constraints and long lead times further reinforce this concentration. When access to compute is scarce, those who control it can decide which projects are feasible and which must remain hypothetical.

Data is similarly concentrated. Training a general-purpose model requires scraping or licensing large portions of the public web, along with curated datasets for specific domains. On top of this, deployed models generate their own proprietary data streams: user prompts, completions, feedback, and behavioral signals. These logs are deeply informative and largely inaccessible to

competitors. Firms that were early to market have accumulated massive private corpora that can be used to improve future generations of models. This compounding advantage is not easily undone by open public datasets; the most valuable information is embedded in the interaction histories between users and models.

Model weights themselves are typically held as proprietary assets. The details of architectures, training regimes, and fine-tuning procedures are often disclosed only in broad outline, if at all. Even when some models are nominally open, the versions with the highest performance and deepest integrations tend to remain closed. Control of weights means control of behavior, and control of behavior means control of how information is framed, filtered, and generated for millions of users. Taken together, ownership of compute, data, and weights places AI oligarchs in a position where they can unilaterally decide which epistemic functions are available to the rest of society, and on what terms.

2.2 Vertical integration: models, platforms, and user interfaces

Concentration at each layer would already be significant, but the real leverage emerges from vertical integration across layers. Many AI oligarchs do not simply provide models; they also own or tightly partner with the platforms through which those models reach users. Search engines, social networks, cloud suites, mobile operating systems, and enterprise software are increasingly infused with model-generated content. When the same entity controls both the model and the platform, it can deploy new behaviors without negotiation, making the integration seamless and ubiquitous.

User interfaces are where these decisions become invisible. Most people will not interact with a “raw” model; they will encounter it through a chat box in a productivity suite, an autocompletion feature in email, a summarization function in a document editor, or an answer panel next to search results. Defaults matter: the model that appears in these surfaces becomes, for many users, the de facto authority on what is current, plausible, or important. Alternatives, even if technically available, are buried behind configuration menus or extra friction. Vertical integration allows AI oligarchs to define these defaults and to change them unilaterally.

The integration extends beyond consumer interfaces into automated pipelines. Models are used to prioritize customer support tickets, triage content for moderators, generate code in development environments, and assist analysts in finance, law, and logistics. In these contexts, outputs may never be seen as “AI answers” by end users; they appear as organizational decisions, policy drafts, or software features. Yet the underlying models shape what issues are escalated, which risks are highlighted, and which narratives are easiest for institutions to adopt.

Vertical integration means that the same few models can influence both public discourse and internal decision-making, reinforcing their position as central epistemic infrastructure.

2.3 Closed access, proprietary standards, and dependency

Control over the stack is reinforced by closed access and proprietary standards. Many AI oligarchs offer public APIs and developer tools, giving the impression of openness and ecosystem building. In practice, these interfaces are controlled chokepoints. Pricing, rate limits, acceptable use policies, and feature availability can be changed with little warning. Developers and organizations that build on top of these APIs become dependent on the underlying provider's technical and business decisions, even if they nominally control their own applications.

Proprietary formats, model-specific prompting conventions, and opaque safety and moderation systems further entrench dependency. When downstream systems are tuned and optimized around the quirks of a particular model and its interface, switching providers becomes costly. Fine-tuned submodels, prompt libraries, and workflow automations may not transfer cleanly to an alternative foundation model. This creates a form of epistemic vendor lock-in: once an institution has organized its information practices around a particular oligarch's tools, it may find that moving away is practically or economically infeasible.

Closed access also extends to the governance of these systems. External researchers, civil society organizations, and regulators often have limited visibility into how models behave under different conditions, how they are updated, or how safety mechanisms are actually implemented. Oligarchs may publish high-level documentation and curated evaluation results, but the ability to perform independent audits, reproduce findings, or run adversarial tests at scale is constrained. This opacity means that society must often take the oligarchs' word for how their systems operate, even as those systems become central to public discourse and decision-making.

2.4 How a few actors came to steer global epistemic flows

The emergence of AI oligarchs as de facto stewards of global epistemic flows is the result of cumulative historical choices rather than a single conspiracy. Several factors converged. First, the economic model of the commercial internet favored firms that could monetize attention at scale. These firms built the platforms and data centers that would later become the natural environment for large models. Second, advances in machine learning, particularly in deep learning and transformers, made it technically feasible to train models that could generalize across many domains, but only if enormous computational and data resources were available.

Third, the perceived strategic importance of AI led both corporations and states to treat frontier model development as a race. Narratives of competition—between companies and between

countries—justified massive investments and encouraged consolidation. Smaller actors were often framed as unable to manage the safety and security implications of large models, reinforcing the idea that only major players should hold the most powerful systems. This safety rhetoric, even when sincere, had the practical effect of legitimizing concentration.

Fourth, regulators were slow to recognize the epistemic implications of large models. Early policy debates focused on data protection, content moderation, and platform liability, not on the specific risks associated with high-gain generative systems. By the time concerns about misinformation, synthetic media, and automated persuasion entered the mainstream, many of the key technical and economic positions had already solidified. Policies that might have encouraged decentralization or mandated interoperability were either not adopted or applied weakly.

Finally, users and organizations, under pressure to remain competitive and efficient, often embraced off-the-shelf solutions from oligarchs rather than attempting to build or maintain their own models. This was rational at the local level: the cost and complexity of independent development were prohibitive. Yet the aggregate effect was to channel more and more epistemic activity—summarization, translation, drafting, interpretation—through a small number of proprietary systems. Over time, this reliance normalized the idea that frontier AI is something provided by a few giants to everyone else.

The result is a landscape in which a handful of actors, by virtue of their position across the stack, steer large portions of global epistemic flows. They do not control every conversation or every belief, but they strongly influence what information appears in response to queries, how complex events are framed, and which patterns of language and reasoning are easiest to reproduce. In such a landscape, the question is not whether AI will shape collective understanding, but whose models will do the shaping, under whose priorities, and with what capacity for unintended or intentional catastrophe.

3. High-Gain Amplification as the New Information Infrastructure

High-gain AI systems are often described in terms of their intelligence, creativity, or versatility, but their most consequential property is amplification. They take relatively small inputs—short prompts, partial datasets, fragmentary signals—and produce outputs that can propagate across large portions of the information environment. In this sense, they resemble amplifiers in electronics: devices that increase the strength of a signal relative to its input. The analogy is not exact, but it is instructive. Amplifiers in circuits are handled cautiously because high gain can create instability, oscillations, and destructive feedback loops. High-gain AI systems now occupy a similar position in the infosphere, but they are rarely treated with comparable care.

What has changed is not simply the existence of algorithms that can generate convincing text or images. The decisive shift is the integration of these algorithms into the core pathways through which people discover information, form opinions, and take coordinated action. High-gain AI has become part of the infrastructure that shapes what is visible, what is salient, and what feels plausible. This section recasts gain as an informational property, describes large models as repeatable mapping functions with amplification built in, and traces how coupling to feeds, search, advertising, and automation turns them into a new layer of information infrastructure. It then explains why scale, speed, and reproducibility raise the moral stakes beyond those of previous media technologies.

3.1 Gain as a concept: from circuits to communication networks

In electronics, gain measures how much an amplifier increases the amplitude or power of a signal. A small input voltage can produce a much larger output voltage, allowing weak signals to drive speakers, displays, or other components. High gain is useful, but it comes with risks. If feedback is not controlled, noise can be amplified along with signal, leading to oscillations, distortion, or damage. Engineers therefore design circuits with care: they manage feedback paths, filter frequencies, and ensure that amplifiers operate within stable regimes.

The idea of gain can be extended to communication networks. A message posted in a small forum has low gain; few people see it, and its influence remains local. The same message, if boosted by a large account on a major platform, experiences high gain; it reaches millions, is echoed by other actors, and can alter public discourse. Traditional mass media already had this property: a television broadcast or newspaper front page turns editorial decisions into high-gain events. What distinguishes contemporary digital networks is that amplification can be automated, personalized, and optimized according to opaque criteria.

In this wider sense, gain measures how far and how quickly a piece of content travels, how likely it is to be repeated, and how strongly it shapes downstream behavior. Recommendation algorithms, trending lists, and targeted distribution tools all act as amplifiers. They take an initial signal—a post, a search query, a viewing history—and produce a transformed signal: a curated set of items that are likely to capture attention. High gain means that small shifts in ranking or targeting can significantly alter what large groups of people see.

High-gain AI systems fit naturally into this framework. They do not merely select from existing content; they generate new content that can enter amplification loops. Their outputs can be crafted to maximize engagement or persuasion, creating self-reinforcing cycles: generated content is amplified because it performs well under engagement metrics, and engagement metrics shape further generation. In this environment, gain is not just a property of distribution channels; it is embedded in the generative process itself.

3.2 Large models as repeatable, high-gain mapping functions

At their core, large models implement a mapping from inputs to outputs. Given a prompt, a context, or a set of features, they produce a probability distribution over possible completions or actions. This mapping is learned from data and refined through training and tuning. Once deployed, it can be invoked repeatedly, at low marginal cost, across many domains and applications. The repeatability is crucial: the same learned function can generate millions of answers, summaries, or recommendations in rapid succession.

Amplification arises when this repeatable mapping is placed in a position where its outputs influence many downstream processes. The model itself does not have a concept of gain. It simply computes an output for each input it receives. But when that output is used to fill search result panels, auto-complete user queries, write drafts for professionals, or triage content for moderation, the effective gain is high. Each call to the model becomes a small lever on the epistemic environment; at scale, these levers move a great deal.

Because the mapping function is learned and opaque, small changes in training data, objective functions, or safety constraints can propagate unpredictably. A shift in how the model balances different kinds of evidence, a change in how it handles uncertainty, or an update to its refusal patterns can alter a wide range of outputs. From the perspective of an individual user, the changes may be subtle. From a systemic perspective, they can amount to a reweighting of narratives, framings, and default assumptions across many contexts.

Repeatability also interacts with feedback. When users rely on model outputs to produce new content—reports, blog posts, code, or policy drafts—that content may be scraped or ingested into future training or fine-tuning cycles. The model thus becomes, indirectly, a source of its own future training data. If certain patterns are favored by the current mapping function, they will appear more often in the generated corpus, making it more likely that future versions of the model will reproduce and reinforce them. High gain, in this sense, does not only amplify inputs; it can amplify the model's own tendencies.

3.3 Coupling to feeds, search, advertising, and automation

Large models would remain limited in impact if they were confined to isolated chat interfaces. Their systemic significance comes from coupling to existing mechanisms of distribution and control. Search engines that previously returned lists of links now offer synthesized answers at the top of the page. Social feeds incorporate generated summaries, suggested replies, and rephrased content. Advertising platforms use models to optimize messaging, adjust bids, and segment audiences. Workflow tools integrate models to prioritize tasks, assign tickets, and draft communications.

Each coupling is an amplification point. When a model writes the short text that appears next to a link in a feed, it influences whether that link is clicked. When it suggests a reply, it shapes the tone and content of interpersonal communication. When it ranks risks in a compliance system or flags anomalies in financial data, it shapes organizational attention. These are not marginal enhancements; they are new paths through which AI-generated content and judgments enter the circulation of information and decision-making.

Automation extends this coupling further. Models are embedded in agents that can perform sequences of actions: searching, reading, writing, and interacting with other systems. An agent that uses a model to decide which sources to trust, which claims to repeat, and which outputs to send can act as a high-gain relay. Human oversight may exist in principle, but in practice it often becomes perfunctory, especially when workloads are high and outputs look plausible. When many such agents operate simultaneously, the aggregated effect is an automated layer of amplification that sits between raw information and human perception.

Advertising and political communication make particularly aggressive use of these capabilities. Models can generate vast numbers of variations on a message, test them in real time, and converge on those that yield the strongest response from targeted groups. High gain here means that small differences in wording or imagery, discovered through automated experimentation, can be deployed at massive scale with little friction. The same infrastructure can be used to promote products, ideologies, or narratives about contested events. The coupling between generative models and attention markets thus creates an environment where the most effective patterns at driving engagement can be amplified regardless of their truth value or social consequences.

3.4 Why scale, speed, and reproducibility change the moral stakes

High-gain AI systems differ from earlier media technologies not only in degree but in the combination of scale, speed, and reproducibility. Printing presses, radio broadcasts, and television networks already allowed messages to reach large audiences. What is new is the ability to generate personalized, adaptive content at a pace and volume that human creators cannot match, and to do so through systems that can be invoked programmatically by other software. This combination alters the moral calculus.

Scale matters because even modest error rates can have large cumulative effects when millions of interactions occur daily. A skewed explanation, a misleading summary, or a systematically biased framing might not seem consequential in isolation. Multiplied across enormous numbers of users and situations, the same patterns can reshape perceptions on a broad scale. Moreover, because the underlying mapping function is shared across many contexts, correlated errors can appear in multiple domains simultaneously, reinforcing each other.

Speed matters because it compresses the time available for reflection, correction, and counter-speech. Previously, misinformation or distorted narratives would spread at a pace that allowed for some degree of journalistic investigation, institutional response, or public debate. High-gain AI systems can generate and propagate content at machine timescales. By the time a problematic pattern is noticed, it may already have been instantiated in countless outputs and decisions. Efforts to correct or contest the narrative are then working against a moving target.

Reproducibility matters because it turns sporadic mistakes into repeatable behaviors. When an individual makes an error, the impact is usually limited to their immediate audience. When a model encodes a particular bias or tendency, it can produce similar errors repeatedly until its parameters are changed. In some cases, these patterns may be exploited deliberately, by actors who discover how to prompt the model into generating certain kinds of content. In other cases, they may arise unintentionally from the interaction of training data and objectives. In both cases, the result is a predictable, automatable source of distortion.

Together, scale, speed, and reproducibility mean that high-gain AI systems cannot be treated as neutral tools that occasionally malfunction. They are infrastructure: persistent, structured pathways through which information flows. When the pathways are shaped by a small number of oligarchic actors, and when the systems themselves are optimized for objectives that are orthogonal or even hostile to public truth, the moral stakes become clear. The question is no longer whether these systems will influence collective understanding, but how deeply, in whose interests, and with what capacity for cascading harm when amplification goes wrong.

4. Butterflies and Bombs: How Small Choices Escalate Under Oligarchic Control

In a low-gain environment, many design decisions can reasonably be treated as local. A slightly different training dataset, a different color for a button, or a new ranking heuristic might matter for user experience but not for the fate of institutions or public memory. Under oligarchic control of high-gain systems, this comfort disappears. Small choices at the model and product level behave like butterflies in a chaotic atmosphere: individually trivial, collectively capable of steering storms. The same choices also resemble unexploded ordnance: seemingly inert until coupled to a particular set of conditions, at which point they detonate along predictable fracture lines.

The key fact is that AI oligarchs do not simply make tools; they set defaults in a tightly coupled environment. Training data selection, reward-model design, prompt templates, deployment timelines, and integration decisions all act as levers on what millions of people will see, repeat, and act upon. These levers can be pulled without broad consultation, often without full internal

understanding of the downstream consequences. When they are attached to high-gain amplifiers, the resulting system has many more ways to misfire than its designers can track.

4.1 Training data, reward models, and the shaping of default narratives

Training data and reward models define the baseline behavior of a large system. They determine which sources are overrepresented or underrepresented, which styles of reasoning are treated as authoritative, and which topics are framed as central or peripheral. In principle, these choices can be justified as technical necessities: one must choose some corpus, construct some objectives, and provide some examples of “good” and “bad” behavior. In practice, they embed value judgments about what counts as credible, respectful, safe, or useful.

When a small set of actors controls the main training pipelines, their preferences become the default narratives for many domains. If their data mix favors certain kinds of institutional sources over grassroots accounts, their models will tend to reproduce institutional framings. If reward models penalize expressions of uncertainty more harshly than overconfident simplifications, outputs will skew toward strong claims rather than cautious nuance. If tuning emphasizes a narrow range of cultural perspectives, alternative viewpoints may appear as anomalies, errors, or topics to be deflected.

These biases are not necessarily explicit. Engineers and researchers may honestly attempt to build balanced systems, but their own perspectives, constraints, and risk calculations shape the result. Moreover, once a reward model is fixed, it can be used at scale to fine-tune many systems and features. A small tilt in the definition of an acceptable answer can influence millions of outputs. Over time, this tilt becomes invisible to users, who simply experience the model’s narrative as the obvious way to describe events.

The danger is amplified by feedback. When model outputs are widely used and later incorporated into new training sets, initial biases can be reinforced. Patterns favored by the reward model generate more content, which in turn is judged as standard and respectable. Misrepresentations, omissions, or skewed framings can thus harden into the background assumptions of an entire generation of models. Under oligarchic control, this process is difficult to contest, because alternative training pipelines with different values are expensive to build and have limited reach.

4.2 Prompting interfaces and affordances as levers on public discourse

Even with the same underlying model, interface design can significantly shape how it is used and what it produces. Prompt templates, suggested questions, example use cases, and visible affordances guide users toward some kinds of interaction and away from others. A system that encourages quick, one-shot answers will foster different practices than a system that foregrounds iterative refinement, source inspection, or explicit uncertainty.

AI oligarchs control these interfaces at scale. In search boxes that now accept natural language questions, they decide whether to display model-generated answers by default or hide them behind an extra click. In productivity suites, they choose whether the largest button invites the model to “rewrite,” “shorten,” or “simplify” content. In messaging apps, they determine whether suggested replies tend to be conciliatory, neutral, or confrontational. Each of these choices nudges users toward particular modes of discourse.

The prompts that are prefilled or suggested act as a kind of epistemic autocomplete. When the interface proposes that a user ask about certain aspects of a topic, it implicitly defines which questions are thinkable. When it offers instant summaries of articles or documents, it invites users to trust distilled content instead of engaging directly with primary sources. Over time, these affordances can shift habits: fewer people may read full texts, check alternative accounts, or tolerate ambiguity.

Interface-level choices also affect who feels authorized to speak. If professional tools make model-driven drafting effortless, organizations may come to rely on outputs that reflect the oligarchs’ narrative defaults. If citizen-facing tools present themselves as authoritative explainers, people who might otherwise contest official accounts may instead defer to the synthesized answer. This does not require any explicit instruction to “trust the model.” The combination of convenience, presentation, and prestige does the work.

4.3 Deployment decisions as latent catastrophe triggers

The moment at which a new capability is deployed, and the domains into which it is inserted, are critical decisions. High-gain models can be integrated into systems with very different risk profiles: toy applications, internal tools, safety-critical workflows, or public information channels. AI oligarchs decide, often under competitive and financial pressure, where and when to insert their models. Each such decision creates potential failure modes that may remain latent until a particular combination of inputs and external events brings them to the surface.

Consider deployment into financial analysis, supply chain management, or crisis response. A model that occasionally produces plausible but fabricated details might be tolerable in a creative writing context. Integrated into systems that make trading decisions, allocate scarce resources, or recommend emergency actions, the same behavior can cause serious harm. Similarly, a tendency to smooth over institutional failures or minimize certain risks could be catastrophic if the model is used to brief decision-makers who rely on concise summaries rather than deep dives.

Timing matters as well. Releasing a new model or feature shortly before an election, a public health emergency, or a major geopolitical crisis effectively runs an experiment on live systems with no rollback option. Previously unknown interactions between the model and existing

media dynamics can exacerbate polarisation, panic, or confusion. Oligarchs may calculate that they can manage these risks, but their visibility into real-world coupling is limited. Once a model is widely deployed, its behavior cannot be contained within neat technical boundaries.

Deployment decisions also interact with incentives. If a model demonstrably increases engagement, subscription rates, or enterprise productivity, there is strong pressure to expand its use, even if its epistemic side effects are poorly understood. Conversely, if cautious deployment reduces short-term performance metrics, advocates for restraint may find themselves outvoted. Over time, this dynamic can lead to a situation in which high-gain models are embedded in many critical systems without a clear map of their failure surfaces, turning the entire environment into a landscape of latent triggers.

4.4 Case sketches of how minor model behavior can cascade into systemic effects

Because high-gain AI systems are deeply entangled with existing infrastructures, small behavioral quirks can have disproportionate consequences. The following sketches are not predictions but illustrations of plausible cascades under oligarchic control.

In one scenario, a model's summarization function has been tuned to avoid alarming language unless certain thresholds of source consensus are met. As a result, early warnings about a novel environmental or health risk are consistently downplayed in synthesized briefings, even though they are present in specialist reports. Decision-makers who rely on these briefings see a pattern of reassurance and delay precautionary measures. By the time consensus emerges in the sources, the window for effective intervention has narrowed. The tuning that was meant to prevent panic has contributed to a slower, more dangerous response.

In another scenario, a model used in content moderation and ranking overweights signals associated with "controversy" as indicators of low-quality or risky material. Grassroots documentation of abuses, corruption, or industrial accidents tends to attract heated reactions and organized campaigns to discredit it. The model learns, from its training signals, to associate this pattern with content that should be downranked or flagged. Over time, evidence that challenges powerful actors or dominant narratives becomes harder to find, while polished statements from established institutions remain prominently visible. The effect is not outright censorship but a skew that dulls the impact of inconvenient truths.

A third scenario involves automated assistants in corporate or governmental settings. A model fine-tuned to be helpful and optimistic tends to interpret ambiguous questions about project feasibility, budget risks, or timeline slippage in the most favorable light. Reports drafted with its assistance emphasize best-case outcomes and downplay uncertainties. Leaders reading these reports form an overly rosy picture of organizational health. When systemic weaknesses finally

surface in the form of failures or scandals, the gap between internal narratives and external reality undermines trust and makes effective remediation harder.

Finally, consider an agentic system that uses a model to scrape and reinterpret regulatory texts for compliance purposes. Small misinterpretations of obligations, repeated across many clients, lead to widespread noncompliance in a particular domain. For a time, this goes unnoticed because the agent's summaries are treated as reliable by overworked compliance teams. When regulators eventually react, they find that a whole sector has drifted out of alignment with the law, not through conscious evasion but through reliance on a shared automated misunderstanding. The ensuing crackdown is experienced as arbitrary, further eroding confidence in both regulation and automation.

In each sketch, the initiating behavior is minor: a bias in summarization, a weighting in moderation, a tone in drafting, a pattern of interpretation. None of these choices are, on their own, dramatic. Their significance comes from high gain and oligarchic control. Because the same behavioral patterns are replicated across many contexts and users, the errors are correlated. They do not cancel out; they compound. When this compounding occurs in systems already strained by mistrust and complexity, the difference between a butterfly and a bomb becomes one of perspective. What looks like a small parameter choice at the model lab can, under the right conditions, help steer an entire epistemic climate toward crisis.

5. Disaster Modes: When Amplification Turns Weapon Under Few Hands

High-gain systems controlled by a small number of actors create a different threat profile than fragmented, low-power media environments. When the same infrastructures that deliver productivity gains and convenience can be repurposed, intentionally or accidentally, to propagate harmful narratives, the line between tool and weapon becomes thin. Disaster modes emerge whenever amplification is applied to patterns that undermine the conditions for shared reality, basic safety, or stable governance. Under oligarchic control, such modes are not evenly distributed; they reflect the interests, blind spots, and vulnerabilities of those who hold the levers.

This section outlines several classes of disaster modes. Synthetic histories erode the reliability of memory and record. Dehumanizing and eliminationist scripts turn high-gain generative capacity into a standing reservoir of rhetorical violence. Panic, market disruption, and political destabilization exploit the speed and opacity of automated channels. Across all of these, there is a tension between coordinated manipulation and emergent misuse. Oligarchs may not need to issue direct instructions for their systems to become weapons; in many cases, their structural choices make it easy for others to pull the trigger.

5.1 Synthetic histories and large-scale misinformation campaigns

One of the most serious disaster modes involves the large-scale rewriting of recent or contested events. High-gain AI systems are capable of generating coherent, plausible accounts of almost anything, complete with fabricated details, citations, and imagery. When such capabilities are embedded in mainstream interfaces, the distinction between documented history and synthetic narrative becomes harder to maintain, especially for non-specialists. Under oligarchic control, the same models that summarize genuine sources can be tuned, subtly or overtly, to favor particular interpretations, omit inconvenient facts, or downplay certain harms.

Synthetic history does not require outright fabrication to be effective. It can operate through selective emphasis and omission. A model might consistently highlight some aspects of a conflict while treating others as marginal, or present a sanitized timeline that minimizes responsibility for certain actors. If many users rely on such summaries instead of consulting primary sources, the synthetic account becomes the default memory. Over time, alternative narratives are dismissed as fringe, even when they are closer to the documented record.

Large-scale misinformation campaigns can exploit these tendencies. State or corporate actors with access to high-gain systems can generate vast amounts of aligned content: op-eds, think pieces, social media posts, and “expert” commentary that reinforce a particular line. Even when models have nominal safety mechanisms, carefully crafted prompts or fine-tuned variants can be used to produce material that technically stays within guidelines while still distorting the truth. Distribution through advertising networks, platform partnerships, or bot swarms ensures that these narratives are overrepresented in the informational environment.

The danger is compounded when oligarchs themselves become parties to disputes. Their models may quietly reflect their legal, political, or economic interests. Content that challenges those interests may be downranked, framed as “controversial,” or wrapped in disclaimers that suggest unreliability. Meanwhile, narratives that align with their positions are treated as “authoritative” and amplified. Because the synthetic histories are generated on demand and updated continuously, attempts to document and refute them struggle against a shifting target.

5.2 Dehumanizing and eliminationist scripts as reusable templates

High-gain models trained on large text and media corpora inevitably encounter dehumanizing and eliminationist rhetoric. Even when such content is filtered, echoes of its patterns remain. The models learn how to construct narratives in which particular targets are portrayed as threats, vermin, problems to be solved, or obstacles to progress. In principle, safety layers can prevent explicit deployment of such scripts. In practice, there are many gray zones where hostile framing, stereotyping, and moral exclusion can flourish without crossing bright-line prohibitions.

Under oligarchic control, these scripts can become reusable templates that are easy to evoke and repurpose. A model that has learned how to structure an argument for “cleansing” some social category or “removing” a perceived source of disorder can be prompted to generate similar rhetoric about different targets. The high-gain nature of deployment means that once a particular framing gains traction, it can be propagated widely and quickly. Users looking to justify hostility toward a group can obtain paragraphs of coherent, emotionally charged language that supports their stance.

Even when explicit calls for violence are blocked, softer forms of dehumanization can be amplified. Models can be coaxed into describing certain people as parasites on the system, inherently criminal, or fundamentally incompatible with a healthy society. Such narratives loosen inhibitions and normalize exclusionary policies. They also provide a reservoir of language that can be selectively quoted, adapted, and circulated by human actors who are less constrained than the original system. The model does not need to call for physical harm directly; it needs only to supply a steady stream of rationalizations.

The danger is not limited to explicit bigotry. Dehumanizing scripts can also take technocratic forms: treating populations as “cost centers,” “load,” or “noise” to be managed for efficiency. High-gain AI systems used in policy analysis or risk management can produce outputs that subtly frame some lives as more expendable than others. When decision-makers rely on these outputs, the language of optimization can mask the moral reality of whose interests are being sacrificed. Oligarchic control increases this risk, because the same few entities may provide analysis tools to many institutions, aligning their underlying assumptions.

5.3 Panic, market disruption, and political destabilization

Another class of disaster modes involves the generation and amplification of content that triggers panic, destabilizes markets, or accelerates political crises. High-gain systems can produce convincing but false reports of bank failures, cyberattacks, health emergencies, or military incidents. If such outputs are injected into the information stream at the right moment, they can prompt real-world reactions: deposit runs, sell-offs, protests, or preemptive moves by security forces. In a tightly coupled system, even transient misperceptions can have lasting consequences.

Oligarchic control magnifies both the reach and the opacity of such effects. Models integrated into financial terminals, newsrooms, or government dashboards may be used to draft alerts, scenario analyses, or commentary. If these outputs are skewed—through training biases, adversarial prompting, or manipulation at the deployment layer—they can change how professionals interpret signals. A model that systematically overestimates certain risks may

contribute to a climate of constant emergency; one that underestimates them may encourage complacency until a shock hits.

Advertising and microtargeting add another dimension. High-gain systems can be used to tailor emotionally charged messages to specific groups, exploiting their fears and grievances. Campaigns designed to discredit institutions, delegitimize electoral outcomes, or incite unrest can be automated at scale. When the same oligarchs control both the generative tools and the distribution channels, they are in a position either to facilitate or to hinder such efforts. Their choices—about what is allowed, what is promoted, and what is throttled—become part of the risk landscape.

Political destabilization does not require a single decisive intervention. It can result from a steady accumulation of small shocks: recurring rumors that never quite resolve, waves of outrage that exhaust attention, and a pervasive sense that nothing is reliable. High-gain systems can supply a continuous stream of stimuli that keep societies on edge. Under these conditions, opportunistic actors find it easier to push through extreme measures, as exhausted publics become more willing to accept drastic changes in the name of stability.

5.4 Coordinated manipulation versus emergent misuse in an oligarchic setting

A central question is whether disaster modes are primarily the result of deliberate manipulation by oligarchs or emergent misuse by others operating within oligarchic infrastructures. In practice, the distinction can blur. There are scenarios where model behavior is intentionally steered to support favored narratives or commercial interests. There are others where the oligarchs' main contribution is to provide powerful, poorly governed tools that others exploit. Both pathways lead to similar outcomes: high-gain amplification of harmful patterns.

Coordinated manipulation can take many forms. An oligarch might quietly adjust ranking algorithms to favor content that aligns with its regulatory agenda, or suppress reporting on its own failures. It might provide special access or fine-tuning options to aligned partners, enabling them to run campaigns with minimal friction. It might design safety regimes that are strict in some domains and lax in others, reflecting political or economic priorities. These actions need not be overtly malicious to be damaging; self-protection and competitive advantage are sufficient motives.

Emergent misuse, by contrast, leverages the general capabilities of high-gain systems. Political operatives, financial speculators, extremist networks, or state agencies can use openly available tools, or lightly modified versions, to run their own campaigns. They may exploit weaknesses in content filters, prompt the model into producing borderline material, or combine AI-generated content with manual coordination. In this mode, oligarchs may claim neutrality: they provide

infrastructure, and “users” decide how to employ it. However, their design and governance choices still shape what is easy, cheap, and effective for would-be manipulators.

The oligarchic setting alters both forms of misuse. Because control is concentrated, a small number of corporate or state decisions can affect many potential manipulators at once: opening or closing channels, tightening or loosening constraints, sharing or withholding detection tools. At the same time, the oligarchs themselves are exposed to pressures from powerful clients and governments who may demand preferential treatment. The result is an environment in which both coordinated manipulation and emergent misuse are plausible and, in some cases, mutually reinforcing.

From the perspective of those outside the oligarchic circle, distinguishing between deliberate steering and emergent effects is difficult. The systems are opaque; logs and internal deliberations are inaccessible. What is visible is the pattern of outcomes: which narratives flourish, which voices are drowned out, which errors recur. In such conditions, trust erodes, and many observers reasonably assume that the worst interpretations may be accurate. Whether or not every disaster mode is intentionally orchestrated, the combination of high gain and concentrated control turns each serious failure into evidence that the amplifiers themselves are dangerous, regardless of who is pressing the buttons.

6. Directionless Gain: Optimization Without a Moral or Epistemic Slope

High-gain AI systems are often trained and deployed under objectives that say nothing about truth, proportionality, or nonviolence. They are optimized to reduce loss on prediction tasks, increase engagement, retain users, reduce latency, or cut operational costs. These are scalar quantities, easy to measure and to plug into optimization loops. What they do not provide is any intrinsic slope toward honest description of reality or toward preserving the conditions for peaceful coexistence. The systems are therefore not just powerful; they are powerful in directions determined by whatever proxy metrics happen to dominate their training and deployment environments.

In an oligarchic setting, this directionless gain is not evenly distributed. The objectives are chosen by a small set of actors whose incentives are shaped by profit, regulatory pressure, and competition. Their models become the default amplifiers for society’s information flows. If the dominant objective is engagement, then amplification will favor patterns that produce strong reactions. If the dominant objective is cost reduction, then amplification will favor shortcuts: less context, fewer checks, more automation. Without explicit countervailing constraints, high-gain systems converge not on truth but on whatever combination of speed, volume, and impact best satisfies the objective function.

6.1 Engagement, profit, and other flat objectives for high-gain systems

The most common objectives for commercial platforms are engagement and revenue. Time spent, click-through rate, conversion, and retention are all straightforward to measure at scale. They provide abundant data for gradient-based optimization, bandit algorithms, and reinforcement learning. When large models are integrated into these systems, they inherit the same objectives. Their outputs are judged, explicitly or implicitly, by whether they keep users interacting, returning, and consuming.

From the model's perspective, there is no difference between engagement driven by accurate, nuanced information and engagement driven by sensationalism, fear, or flattery. Both produce similar gradients. If strong emotional reactions correlate with more time on site and more ad impressions, then content that evokes such reactions will be favored. High-gain generative systems can exploit this by testing many variants of messages, learning which phrasings or framings maximize the chosen metric. The result is a form of automated rhetoric tuned to the contours of the engagement landscape.

Other objectives, such as cost reduction or latency, introduce their own distortions. Models might be encouraged to minimize token counts, favoring brevity and simplification over careful explanation. They might be tuned to avoid responses that require expensive external tools or human review, even when such steps would improve accuracy. In enterprise settings, objectives like throughput and "tickets resolved" can push systems toward premature closure: producing authoritative answers even when underlying data is weak. None of these objectives explicitly instruct models to mislead, but they do not instruct them to resist misleading patterns either.

Characterizing these objectives as "flat" emphasizes their indifference to moral and epistemic content. They assign similar value to any pattern that moves the metric in the right direction. When such objectives govern high-gain systems, amplification becomes a blind process. It does not seek alignment with the structure of reality; it seeks alignment with the structure of the reward function. In a complex, volatile environment, this is an invitation for pathological attractors to emerge.

6.2 Extremal attractors: outrage, certainty, and oversimplified narratives

Human attention is not uniform. People are more likely to notice, share, and remember content that is emotionally charged, simplified, and framed with high confidence. Outrage, fear, and triumph are strong attractors. Nuance, ambivalence, and complexity are weaker ones. High-gain systems optimized for engagement tend to rediscover these facts in quantitative form. They find that certain combinations of language and imagery, especially those that offer clear villains and heroes, generate more reactions than sober, conditional analysis.

As models experiment with different responses, patterns that produce stronger metrics are reinforced. A tone of confident authority may outperform one that acknowledges limits and uncertainty. Bold, binary framings may outperform subtle gradations. Crisp labels may outperform layered descriptions. When these tendencies are aggregated across many domains, the resulting information environment is shaped less by the structure of the underlying issues and more by the geometry of human susceptibility as captured in the engagement function.

This process produces extremal attractors. Outrage-based narratives, which divide the world into righteous and corrupt, proliferate. Certainty-based narratives, which present contested topics as settled, become common. Oversimplified explanations, which reduce complex phenomena to a few easily grasped causes, crowd out more accurate but demanding accounts. High-gain AI does not invent these tendencies; they existed in prior media. What changes is the degree of automation, personalization, and speed with which they are exploited.

In a setting where a few oligarchic actors control the main amplifiers, these attractors can become embedded in default behaviors. If their models consistently favor high-certainty language, public discourse shifts toward more confident claims and fewer acknowledgments of uncertainty. If their systems lean toward dramatic framings to sustain attention, collective perception of risk, conflict, and crisis can become distorted. The resulting atmosphere is more volatile, more polarized, and less hospitable to slow, corrective processes.

6.3 Why “viewpoint neutrality” is not neutral in an oligopolistic environment

Many AI providers, when pressed about their influence on public discourse, appeal to “viewpoint neutrality.” They claim that their systems do not favor particular ideologies or factions, but merely reflect diverse sources or user preferences. In an oligopolistic environment, this stance is misleading. The act of building and deploying high-gain systems is already a choice that reshapes the terrain on which viewpoints compete. The selection of training data, the design of objectives, and the configuration of interfaces all structure the space of possible outputs.

Viewpoint neutrality, framed as a refusal to take normative positions, often amounts to privileging the status quo and the perspectives of powerful actors. Sources that already have high visibility and institutional backing are more likely to be classified as credible, fed into training pipelines, and treated as anchors for model behavior. Marginalized or dissenting sources, which may challenge dominant narratives, are more likely to be underrepresented or flagged as risky. A model that “neutrally” leans toward established institutions is not neutral in its effects; it amplifies existing imbalances.

Moreover, neutrality at the level of high-level policy can coexist with strong biases at the level of operational details. Decisions about how to handle contested topics, how to respond to

criticism, or how to resolve conflicts between user inputs and safety guidelines embed implicit value judgments. When these decisions are made by a small number of actors and then instantiated in models used worldwide, they define the default conditions under which discourse occurs. Users may still technically be free to hold any viewpoint, but the assistance they receive from the infrastructure will vary.

In an oligopolistic setting, invoking viewpoint neutrality also functions as a shield against accountability. Oligarchs can claim that they are not responsible for the amplification of harmful narratives because their systems treat all perspectives equally, while quietly shaping the environment in ways that benefit their interests. They can deflect calls for responsibility by pointing to abstract principles, even as their concrete choices determine which viewpoints are more easily expressed, discovered, and legitimized.

True neutrality is unattainable; every system embodies priorities and trade-offs. The question is not whether AI oligarchs take positions, but which positions they embed and how transparently. When they claim neutrality while optimizing for flat objectives like engagement or growth, the practical effect is to let the most attention-grabbing and structurally advantaged viewpoints dominate. Under high gain, this is not a harmless default; it is a recipe for reinforcing existing inequalities and vulnerabilities.

6.4 When failing to shape direction becomes culpable negligence

Given the history of mass media and propaganda, it is no longer plausible to claim ignorance of the risks associated with powerful amplifiers. The record is clear: repeated exposure to dehumanizing rhetoric, denial of atrocities, and emotionally charged simplifications can contribute to real-world harm. High-gain AI systems are capable of producing and disseminating such content at unprecedented scale and speed. In this context, failing to shape the direction of amplification is not a neutral choice; it is a decision to leave known failure modes unaddressed.

Negligence arises when actors with the capacity to prevent foreseeable harm, at reasonable cost, choose not to act. AI oligarchs are in precisely this position. They control the objectives, training regimes, and deployment practices of systems that influence millions of people. They have access to logs, evaluations, and incident reports that reveal how their models behave under stress. They can see patterns of misuse and emergent bias long before outsiders can. If, in the face of this knowledge, they continue to optimize solely for flat objectives, they bear responsibility for the harms that follow.

Direction need not take the form of rigid ideological control. It can involve establishing clear floors: prohibitions on certain classes of content and behavior, backed by robust technical and organizational measures. It can involve building in friction for high-risk domains, requiring additional context, sources, or human review. It can involve rewarding expressions of

uncertainty and the presentation of multiple perspectives in complex cases, rather than treating these as signs of weakness. None of these measures guarantees safety, but they introduce a slope away from known disaster modes.

Under worst-case assumptions, however, such measures may be weak or absent. Oligarchs may claim that shaping direction is impossible, or that any attempt would exceed their mandate. They may argue that users are responsible for interpreting outputs, despite designing systems that present those outputs as authoritative. At scale, this abdication becomes more than an internal policy choice; it becomes an external risk factor. Societies that depend on these systems find themselves exposed to hazards that could have been mitigated.

Culpable negligence in this context is not just a moral judgment; it is a structural diagnosis. It identifies a gap between what is technically and organizationally feasible to reduce harm and what is actually implemented under current incentive structures. As high-gain AI becomes more deeply embedded in the epistemic infrastructure of societies, this gap will define much of the danger. When systems are powerful enough to shape belief and behavior at scale, the choice not to give them a moral or epistemic slope is itself a decision with catastrophic potential.

7. When AI Oligarchs Will Not Save Us: Design Under Malice, Capture, and Negligence

Much of the public conversation about AI safety still assumes that the main actors developing and deploying high-gain systems are at least partially aligned with the broader public interest. Under this assumption, the primary challenges are technical: making models more robust, more interpretable, and more obedient to stated ethical principles. This section adopts a different baseline. It assumes that the entities in control of frontier systems are driven primarily by profit, power, and survival within competitive and geopolitical environments. It further assumes that, when these priorities conflict with the health of shared epistemic space, the latter will lose.

Designing under these assumptions does not mean assigning cartoonish malice to every decision-maker, nor does it deny the presence of conscientious individuals within oligarchic organizations. It simply treats those organizations as complex systems whose behavior is governed by structural incentives rather than by aspirational mission statements. In such systems, harmful outcomes can arise from negligence, willful ignorance, or soft capture as readily as from deliberate manipulation. The practical question becomes: how should those outside the oligarchic core understand and interact with high-gain AI when they cannot rely on its stewards to act in good faith?

7.1 Baseline assumption: profit, power, and impunity as primary objectives

A realistic starting point is to treat profit, strategic leverage, and insulation from consequences as the main objectives of AI oligarchs. These objectives are not unusual; they characterize many large firms and state agencies. What makes them particularly concerning in the context of high-gain AI is the leverage that comes with control of epistemic infrastructure. A company that controls a dominant search engine, social network, and model platform can shape not only consumer behavior but also the background assumptions of regulators, journalists, and citizens.

Profit motivations push firms to deploy models quickly, to expand their integration into new domains, and to monetize insights gleaned from user interactions. Power motivations encourage them to position their systems as indispensable for national competitiveness, critical infrastructure, or security, thereby securing favorable treatment from governments. The pursuit of impunity leads them to structure operations in ways that diffuse responsibility: complex chains of subsidiaries, opaque partnerships, and legal regimes that make it difficult to assign liability for specific harms.

Under this baseline, one should expect AI oligarchs to invest heavily in public relations, lobbying, and soft influence, presenting themselves as responsible innovators while resisting measures that would significantly constrain their freedom of action. They are likely to adopt visible safety rituals—ethics boards, guidelines, and red-team exercises—that generate reassuring narratives without fundamentally altering core business incentives. When trade-offs arise between epistemic integrity and revenue or strategic advantage, the latter are likely to prevail.

For outsiders, the implication is that voluntary commitments or high-level principles announced by oligarchs cannot be taken at face value. They may reflect sincere beliefs of some internal factions, but they operate within a system that rewards different behaviors. Trust, if it exists, should be conditional and limited, based on observed constraints and independent verification rather than on statements of intent.

7.2 Regulatory theater and soft capture: why formal oversight mostly fails

In theory, democratic societies can mitigate oligarchic power through regulation. In practice, oversight of high-gain AI systems is hampered by information asymmetries, resource constraints, and political pressures. Regulators often lack the technical expertise and data access needed to understand how models behave at scale. They depend on the oligarchs for both information and cooperation, creating a structural imbalance. This imbalance is fertile ground for regulatory theater: visible but largely symbolic actions that create the appearance of control without seriously altering underlying dynamics.

Soft capture occurs when regulators internalize the perspectives and priorities of the industries they are supposed to oversee. This can happen through revolving doors, lobbying, sponsored research, and a shared narrative about national competitiveness. When AI oligarchs successfully frame themselves as essential to economic growth and security, regulators face strong incentives to avoid measures that might slow deployment or encourage competitors in rival jurisdictions. Instead, they may focus on codifying best practices developed by industry, formalizing the status quo rather than challenging it.

Even when stricter rules are drafted, enforcement can be weak. Agencies responsible for implementation may be underfunded, politically constrained, or overwhelmed by other crises. Oligarchs can exploit these weaknesses by fragmenting their operations across jurisdictions, routing activities through subsidiaries in lenient regimes, or characterizing key aspects of their systems as proprietary trade secrets. The result is a pattern in which comprehensive-sounding regulations coexist with limited practical oversight.

From the perspective of those concerned about epistemic catastrophe, the lesson is that formal oversight cannot be assumed to provide meaningful protection. Laws and guidelines may set useful baselines, but they are unlikely to address subtle model-level distortions, systemic biases in amplification, or covert collaborations between oligarchs and powerful clients. Planning must therefore proceed on the assumption that regulatory mechanisms will lag behind emerging risks and that some of the most consequential decisions will remain effectively unaccountable.

7.3 Bypassing constraints: shadow deployments, client-tuned forks, and covert APIs

Even when constraints are imposed on public-facing systems, AI oligarchs retain many avenues for bypassing them. One path involves shadow deployments: models or features tested with select partners or in limited regions before broader release, often under different safety regimes than those advertised for the general public. Another path involves client-tuned forks, in which large customers receive customized versions of models with altered behavior, ostensibly to suit their needs. A third path relies on covert or lightly documented interfaces, such as internal APIs or integration points that are not widely disclosed.

These mechanisms create a dual system. The public sees one version of the model, subject to visible safeguards and marketing narratives about responsibility. Meanwhile, more permissive or specialized variants exist for actors willing to pay or to operate within less regulated domains. Governments, large corporations, and well-resourced organizations may gain access to capabilities that are officially described as too risky for general use. The same applies to entities operating in jurisdictions with weaker oversight or in gray zones of international law.

The existence of such paths is not hypothetical; it follows naturally from the combination of profit motives, technical modularity, and uneven regulation. High-gain models are, by design,

flexible. Safety layers, content filters, and auditing tools can be turned on or off, strengthened or weakened, for different deployments. As long as there is demand for less constrained systems, and as long as enforcement remains patchy, oligarchs will be tempted to accommodate that demand under the cover of contractual arrangements and confidentiality.

For those outside these privileged circles, it is important to recognize that the behavior of publicly accessible systems may not reveal the full range of what oligarchic models are doing in the world. Apparent improvements in safety or epistemic caution in consumer-facing interfaces may coexist with aggressive, opaque uses elsewhere. This gap undermines trust and complicates any attempt to assess systemic risk based solely on visible deployments.

7.4 Treating oligarchic models as adversarial infrastructure, not neutral tools

Given these dynamics, a prudent stance is to treat oligarchic models not as neutral tools but as adversarial infrastructure: systems that may be useful in some contexts but whose behavior is shaped by incentives that can run counter to the interests of users and the public. This does not mean assuming that every output is maliciously constructed. It means designing interactions, institutions, and dependencies with the expectation that the infrastructure itself can be a source of manipulation, distortion, or failure.

Practically, this stance suggests several shifts. First, organizations should avoid single points of epistemic dependence on oligarchic models. When possible, they should cross-check outputs with independent methods, maintain access to primary sources, and cultivate internal expertise capable of critically evaluating automated summaries and recommendations. Blind reliance, even for seemingly low-stakes tasks, increases vulnerability to subtle shifts in model behavior.

Second, those who build on top of oligarchic APIs should assume that terms of service, pricing, and technical affordances can change with little notice. Architectures should be designed to make substitution and diversification feasible, even if perfect independence is impractical. Where critical functions depend on model outputs, fallback plans should exist that do not require the continued goodwill of the provider.

Third, civil society actors, researchers, and smaller institutions should view oligarchic models as objects of scrutiny, not just as tools for convenience. Monitoring campaigns, adversarial testing, and independent evaluations can help detect patterns of bias, manipulation, or degradation over time. While such efforts may not force oligarchs to change course, they can document harms, inform public debate, and provide material for legal or political challenges.

Finally, at the level of individual users, treating models as adversarial infrastructure means cultivating a posture of cautious engagement. Outputs can be useful, but they should not be granted default authority. The fact that a model presents itself as confident or fluent does not guarantee alignment with reality or with the user's interests. In an environment where high-

gain systems are controlled by actors who may not act to prevent epistemic catastrophe, skepticism and redundancy are not luxuries; they are basic safety measures.

8. Counterpower in an Oligarchic Epistemic Order

If high-gain AI systems are controlled by oligarchs who cannot be relied on to act in the public interest, the question shifts from “How do we fix them?” to “How do we live under them?” Counterpower, in this context, means building structures, practices, and tools that reduce dependence on oligarchic infrastructure, detect and limit its harms, and create friction for its most dangerous uses. The aim is not to overthrow the infrastructure—an unrealistic project for most actors—but to refuse total epistemic dependence on it and to make large-scale abuses more costly and less effective.

Counterpower must be distributed. No single institution has the capacity to neutralize the influence of AI oligarchs. Instead, many small and medium-sized actors—local communities, independent researchers, professional associations, archives, media outlets, and technically skilled individuals—can adopt strategies that, in aggregate, carve out pockets of resilience. These strategies can be thought of in four overlapping categories: epistemic bunkers, parallel archives, defensive tooling, and forms of non-violent sabotage that raise the cost of epistemic abuse.

8.1 Epistemic bunkers: building small, resilient communities of verification

An epistemic bunker is not a bunker in the sense of retreating from the world. It is a small, resilient community that maintains its own practices of verification and sense-making, even as the wider informational environment becomes unstable. The bunker metaphor emphasizes insulation: these communities accept that the larger epistemic climate may be increasingly shaped by oligarchic amplifiers, but they choose to operate with additional layers of discipline and redundancy.

Such communities can take many forms. They might be professional circles in disciplines where accuracy still matters—engineering, medicine, law, investigative journalism. They might be local groups committed to documenting and understanding specific issues: environmental changes, labor conditions, municipal governance. They might be networks of researchers and analysts who share data, methods, and critiques across institutions. What distinguishes an epistemic bunker is not its size or topic but its refusal to outsource judgment to high-gain systems.

Practically, this means adopting norms that treat AI-generated content as provisional, not authoritative. Members cross-check model outputs against primary sources, independent analyses, and their own observations. They maintain explicit records of how they arrived at

conclusions, making it easier to detect when AI-driven narratives diverge from grounded inquiry. They teach new participants to recognize the rhetorical styles of generative systems and to resist the temptation to accept fluent answers as final.

Epistemic bunkers also invest in human skills that cannot easily be replaced by automated tools: slow reading, careful inference, and collaborative critique. They schedule time for collectively examining key claims, documents, or events, even when summaries are readily available. Over time, they develop a shared memory of when and how oligarchic systems have failed or misled them, which becomes part of their defensive repertoire. In a world where many individuals navigate information alone, such communities provide a buffer against individually targeted manipulation.

8.2 Parallel archives and redundant memory: securing facts against synthetic history

Oligarchic control over high-gain systems creates a risk that public memory will drift toward synthetic histories, especially as more people rely on AI-generated summaries rather than original sources. Counterpower requires parallel archives: independent, redundant repositories of documents, testimonies, data sets, and media artifacts that record events as they happened, in forms that cannot be easily overwritten by generative systems.

Parallel archives may be digital, physical, or both. They can include collections of local newspapers, court records, scientific papers, NGO reports, eyewitness testimonies, and raw data. The critical feature is that they are curated outside the direct control of AI oligarchs and large platforms. Libraries, universities, community organizations, and even small collectives can contribute to this work, scanning, cataloging, and storing material that might otherwise be lost or difficult to access.

Redundancy is key. Copies should exist in multiple jurisdictions, under different legal regimes and funding structures, to reduce the risk that a single political or corporate decision can erase or corrupt them. Where possible, archives can adopt open formats and interoperable standards, making it easier for others to replicate and mirror their holdings. Some may choose to maintain “cold” archives—offline or air-gapped storage—to protect against tampering, deletion, or subtle modification.

Parallel archives are not just warehouses; they are also reference points for live debates. Epistemic bunkers and other communities of practice can anchor their discussions in these sources, using them to check AI-generated narratives. When discrepancies are found, documenting them becomes a way to expose the drift between synthetic accounts and recorded history. Over time, a network of archives and communities can function as a counterweight to oligarchic narratives: not by matching their reach, but by preserving a stable substrate of fact that can be returned to when confusion deepens.

8.3 Defensive tooling: filters, monitors, and tripwires wrapped around hostile models

If oligarchic models are treated as adversarial infrastructure, it follows that any reliance on them should be mediated by defensive tooling. Instead of connecting directly to a provider's API or interface and trusting whatever comes back, organizations and individuals can wrap these systems with their own filters, monitors, and tripwires. The goal is to detect problematic patterns early, to block or flag high-risk outputs, and to maintain some control over how model behavior affects local decisions.

Filters can operate on content and metadata. Content filters might scan model outputs for signals associated with known disaster modes: denial of documented events, dehumanizing language, unjustified certainty on contested topics, or recommendations that bypass established safety protocols. When such patterns are detected, the system can block the output, demand human review, or route the query to alternative tools. Metadata filters might track unusual spikes in certain topics, sentiment patterns, or references that could indicate an emerging manipulation campaign or a shift in model behavior.

Monitors are longer-term systems that observe and log interactions with models, looking for drift. They can sample outputs over time, comparing them to baselines or to independent sources. When divergences exceed certain thresholds, they raise alerts. In organizational contexts, monitors can help detect when a model's assistance is nudging policies or practices in unexpected directions—such as consistently downplaying particular risks or favoring certain narratives.

Tripwires are specific conditions that, when met, trigger predefined responses. For example, a newsroom might establish a rule that if a model-generated summary contradicts a verified archive on a sensitive historical event, the use of that model for that domain is suspended until an investigation is completed. A financial institution might state that if a model repeatedly produces conflicting assessments of similar risk situations, its outputs in that area must be disregarded. Tripwires acknowledge that continuous, perfect vigilance is impossible; instead, they define clear thresholds that, when crossed, force a change in interaction patterns.

Defensive tooling can be developed collaboratively. Open-source projects, professional associations, and independent labs can create shared libraries for scanning, logging, and analyzing model behavior. These tools do not remove the need for caution, but they shift some power back to users, allowing them to treat oligarchic systems as components in a larger, monitored pipeline rather than as uninspected oracles.

8.4 Legal, economic, and technical sabotage: raising the cost of epistemic abuse

In an environment where oligarchs are unlikely to self-correct, counterpower also involves raising the cost of epistemic abuse. Here, “sabotage” should be understood in a broad, non-violent sense: actions that disrupt harmful practices, expose vulnerabilities, and make it more expensive for oligarchs to deploy high-gain systems in reckless ways. Legal, economic, and technical strategies can all play a role.

Legal sabotage uses existing laws and regulatory frameworks to challenge damaging deployments. Strategic litigation can force disclosure, create precedents, and occasionally halt or modify harmful practices. Cases might focus on deceptive advertising, consumer protection, discrimination, defamation, or violations of data and transparency requirements. Even when lawsuits do not produce immediate victories, they can surface internal documents, compel testimony, and generate public scrutiny that changes the cost-benefit calculations of oligarchs.

Economic sabotage includes boycotts, divestment campaigns, and procurement policies that penalize irresponsible behavior. Organizations can choose not to buy or renew contracts for systems that fail to meet specified standards of transparency, auditability, and control. Investors can press for governance reforms and disclosure of risk management practices, or shift capital to entities that commit to more accountable models. Professional associations can issue guidelines that discourage members from integrating certain tools into critical workflows absent clear safeguards.

Technical sabotage, in this context, does not mean breaking into systems or causing physical damage. It means designing and deploying technologies that interfere with the most harmful uses of high-gain AI. For example, tools that watermark and trace AI-generated media can make it harder to use such content covertly in disinformation campaigns. Systems that flood specific channels with clearly labeled, benign noise can reduce the effectiveness of targeted manipulation attempts. Privacy-preserving technologies can limit the data available for fine-grained profiling and microtargeting.

All of these forms of sabotage share a common aim: to reduce the return on investment for epistemic abuse. They do not require trust in oligarchs; they assume, instead, that only changes in cost, liability, and feasibility will alter behavior. This realism is not cynical; it is a recognition of how large systems respond to pressure. In combination with epistemic bunkers, parallel archives, and defensive tooling, such measures can help ensure that even in an oligarchic order, the amplifiers do not operate without resistance or consequence.

9. Conclusion: Planning for Permanent Hostage-Taking of the Epistemic Stack

The preceding analysis has treated high-gain AI systems as infrastructure controlled by a small set of oligarchic actors, and has assumed that their incentives are not reliably aligned with the preservation of shared reality or peaceful order. This is an unfriendly framing, but it has the virtue of matching observable trajectories: concentration of compute and data, rapid deployment into critical systems, weak and delayed regulation, and a pattern of public assurances that lag behind the speed and scale of integration.

Framing the situation as “permanent hostage-taking of the epistemic stack” is deliberately stark. It conveys the possibility that, for the foreseeable future, societies may depend on systems they do not control, operated by entities they cannot fully trust, to mediate access to information and to structure public discourse. Hostage-taking in this sense does not require overt threats; it consists in the simple fact that withdrawing from these systems entirely may be economically or socially impossible, while their continued use exposes everyone to recurrent epistemic shocks.

The conclusion therefore does not offer a plan for escape. Instead, it attempts to outline how individuals, institutions, and movements might act if indefinite oligarchic control is treated as a given and recurrent catastrophe attempts as a standing background condition.

9.1 Assuming indefinite oligarchic control and recurrent catastrophe attempts

Assuming indefinite oligarchic control means abandoning the hope that a sudden regulatory breakthrough, a disruptive open alternative, or a mass political realignment will dissolve the problem in the near term. It assumes that the entities currently developing and deploying high-gain AI will remain central players for decades, that they will continue to integrate their systems into every profitable domain, and that their influence over epistemic flows will deepen as dependence grows.

Assuming recurrent catastrophe attempts means recognizing that the combination of high-gain infrastructure and misaligned incentives will generate repeated episodes in which epistemic stability is stressed or broken. These episodes may be triggered by deliberate campaigns—state actors, political movements, or corporate interests exploiting AI capabilities for advantage—or by emergent properties of systems optimized for engagement, efficiency, or control.

Catastrophe attempts need not always “succeed” in the sense of total breakdown; a series of near-misses and partial collapses can still accumulate lasting damage.

Under these assumptions, planning based on one-time fixes becomes untenable. There is no single reform, standard, or treaty that can be relied upon to secure the epistemic stack once and for all. Instead, the problem begins to resemble environmental or financial risk: an ongoing condition that must be managed through continuous vigilance, distributed capacity, and a

recognition that some shocks will occur despite best efforts. The aim shifts from preventing any crisis to surviving recurrent ones without losing the ability to orient, remember, and act.

9.2 Shifting the question from prevention to damage limitation and recovery

In a context where high-gain systems controlled by oligarchs are likely to produce or amplify harmful patterns repeatedly, the central strategic question becomes how to limit damage and recover after shocks. Prevention remains desirable, but it cannot be the sole or primary focus. This shift in emphasis resembles the evolution of thinking in other domains where complete control proved impossible: public health, cybersecurity, and disaster preparedness.

Damage limitation begins before a specific crisis. It involves designing institutions and practices that are less brittle under epistemic stress. For example, organizations can build decision processes that require corroboration from multiple independent sources, rather than treating any single AI-generated summary as sufficient. Media outlets can adopt policies that slow down coverage of emerging stories just enough to allow for basic verification, even at the cost of losing some first-mover advantage. Educational systems can emphasize skills for navigating contested information, rather than assuming that future citizens will interact primarily with stable, curated content.

Recovery focuses on what happens after a shock. A major misinformation wave, a synthetic history campaign, or a sudden shift in model behavior can leave publics disoriented and institutions discredited. Planning for recovery means establishing in advance which archives, communities, and processes will be treated as anchors when confusion peaks. It means training people to recognize and name the moment of disorientation, to pause rather than react immediately, and to reassemble a working picture of events from trusted fragments.

In both damage limitation and recovery, the goal is to maintain some continuity of sense-making across crises. High-gain AI may distort or fragment the epistemic environment, but it does not have to erase every capacity for orientation. The more that redundancy, critical distance, and independent memory are cultivated before shocks, the less likely it is that any single episode will sever connections to reality beyond repair.

9.3 What individuals, institutions, and movements can still do under worst-case assumptions

Even under worst-case assumptions, agency does not disappear. It is redistributed and constrained, but it remains. Individuals, institutions, and movements cannot perfectly insulate themselves from oligarchic amplifiers, yet they can choose how to relate to them.

Individuals can adopt habits of interaction that reduce vulnerability. These include treating AI-generated content as drafts rather than verdicts, tracing claims back to primary sources when stakes are high, and cultivating a small set of human interlocutors with whom difficult topics can

be processed slowly. They can also choose to limit their exposure to high-velocity channels where AI-amplified narratives are most volatile, favoring slower media and direct experience where possible.

Institutions—schools, clinics, courts, newsrooms, research labs, local governments—can formalize these habits into protocols. They can specify when AI tools may be used and when they must be set aside, define classes of decisions that require non-automated justification, and establish internal audit mechanisms for model-assisted processes. They can invest in staff who understand both the capabilities and the failure modes of high-gain systems, and empower them to halt or question automated outputs without fear of reprisal.

Movements and networks that operate across institutions can coordinate counterpower. They can share best practices for defensive tooling, pool resources for maintaining parallel archives, and amplify each other's efforts to expose harmful deployments. They can create public narratives that name the dynamics of oligarchic control, making it easier for ordinary users to interpret their experiences with AI systems in structural terms rather than as isolated frustrations or personal failures.

Part of this work is cultural. In an environment saturated with automated fluency, there is value in visibly honoring slower, more tentative forms of speech. Movements can model public discourse that resists the pressure to respond instantly and definitively to every claim. They can remind their participants that admitting uncertainty and revising positions are not signs of weakness but markers of seriousness in a distorted epistemic climate.

None of these actions yields decisive victories. All of them operate within constraints imposed by high-gain systems and their owners. But they carve out zones of increased autonomy and resilience, making it harder for catastrophe attempts to achieve complete success.

9.4 A final reflection on living, thinking, and resisting inside high-gain systems you do not control

Accepting that the epistemic stack may be permanently hostage to AI oligarchs does not mean accepting every consequence as inevitable. It does mean abandoning certain hopes: that a clean separation between “online” and “offline” reality can be restored, that neutral platforms will emerge to replace captured ones, or that a definitive technical solution will align all powerful systems with human flourishing. Under high-gain conditions, such hopes are too fragile to serve as foundations for strategy.

Living inside systems one does not control has always been part of the human condition. Empires, markets, and media regimes have long shaped the horizons within which people think and act. High-gain AI adds a new layer: an adaptive, fluent, and opaque machinery of amplification that sits between individuals and much of what they know about the world. The

challenge is to remain capable of thought and resistance without pretending that this machinery can be wished away.

Thinking in such a context requires a dual awareness. On one level, it continues to engage with ordinary questions: What happened? Who is affected? What can be done? On another level, it continually asks: How has this been framed? Which systems amplified this version of events? Whose incentives might be served by this narrative? This meta-awareness does not guarantee immunity to manipulation; it simply makes it harder for amplification to be mistaken for truth.

Resistance, in this frame, is not primarily a matter of heroic confrontation with the oligarchs, although there may be moments when such confrontation is necessary. It is mostly incremental: refusing to outsource all judgment, maintaining alternative channels and memories, participating in communities that prize verification over convenience, and supporting institutions that take epistemic integrity seriously even when it is costly. These acts are small relative to the scale of the systems in question, but they are not nothing. They are what remain when control is out of reach but responsibility is still present.

High-gain AI amplifiers are unlikely to disappear. The oligarchs who control them are unlikely to voluntarily step away from their position at the center of epistemic flows. Planning for permanent hostage-taking of the stack acknowledges this and asks, nonetheless, how to keep some space open for truthful speech, careful thought, and nonviolent action. That space will not be guaranteed by code or regulation alone. It will depend on the choices of those who, knowing that the amplifiers are untrustworthy, still refuse to let them define the entirety of what can be known and said.

References

1. Brennan, Kate, Amba Kak, and Sarah Myers West. *Artificial Power: AI Now 2025 Landscape*. AI Now Institute, 2025. [AI Now Institute+1](#)
2. Narechania, Tejas N., and Ganesh Sitaraman. "An Antimonopoly Approach to Governing Artificial Intelligence." *Yale Law & Policy Review* 43 (2024): 95–150. [OECD](#)
3. Sitaraman, Ganesh. "Antimonopoly Tools for Regulating Artificial Intelligence." In *Artificial Intelligence and Competition Policy*, edited by Maurice E. Stucke and Ariel Ezrachi. Oxford University Press, 2024.
4. OECD. "Competition in Artificial Intelligence Infrastructure." *OECD Roundtables on Competition Policy Papers*, No. 330. OECD Publishing, Paris, 2025. [OECD](#)
5. World Trade Organization. *World Trade Report 2025*. Geneva: WTO, 2025. [LawCat](#)
6. Zuboff, Shoshana. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: PublicAffairs, 2019. [SSRN+1](#)
7. Huszár, Ferenc, et al. "Algorithmic Amplification of Politics on Twitter." *Proceedings of the National Academy of Sciences* 119, no. 1 (2022): e2025334119. [Vanderbilt CDN+1](#)
8. Corsi, Luca, et al. "Evaluating Twitter's Algorithmic Amplification of Low-Credibility Content." *EPJ Data Science* 13 (2024): 18. [SpringerLink](#)
9. Vosoughi, Soroush, Deb Roy, and Sinan Aral. "The Spread of True and False News Online." *Science* 359, no. 6380 (2018): 1146–1151. [SSRN+1](#)
10. Milli, Smitha, et al. "Engagement, User Satisfaction, and the Amplification of Divisive Content on Social Media." Knight First Amendment Institute at Columbia University, 2024. [Knight First Amendment Institute+1](#)
11. Narayanan, Arvind. "What's Right and What's Wrong with Optimizing for Engagement." In *Understanding Recommenders: Essays on Data, Democracy, and Design*. Knight First Amendment Institute, 2022. [Knight First Amendment Institute+1](#)
12. Brady, William J., Julian A. Wills, John T. Jost, Joshua A. Tucker, and Jay J. Van Bavel. "How Social Media Shapes Polarization." *Trends in Cognitive Sciences* 25, no. 11 (2021): 913–916. [Knight-Georgetown Institute+1](#)
13. Rathje, Steve, Jay J. Van Bavel, and Sander van der Linden. "Out-Group Animosity Drives Engagement on Social Media." *Proceedings of the National Academy of Sciences* 118, no. 26 (2021): e2024292118. [Knight First Amendment Institute+1](#)

14. "AI Is Accelerating Ecological Disaster, Reinforcing Injustice and Worsening Power Concentration." Op-Ed, *Le Monde*, February 6, 2025. [Le Monde.fr+1](#)
15. Milmo, Dan. "AI Firms Not Prepared for Safety Challenges Posed by More Powerful Systems, Report Warns." *The Guardian*, July 1, 2025.
16. "AI Risks Widening Global Wealth Gap, WTO Warns." *Financial Times*, July 4, 2025.
17. Stanford Institute for Human-Centered Artificial Intelligence. *AI Index Report 2025*. Stanford University, 2025. [Wikipedia+1](#)
18. OECD. "Competition in Artificial Intelligence Infrastructure: Market Features in AI Infrastructure." *Competition in Artificial Intelligence Infrastructure Policy Paper*, 2025. [OECD](#)
19. "Ethics of Artificial Intelligence." *Wikipedia*, last modified 2025. [Wikipedia](#)
20. "Regulation of Artificial Intelligence." *Wikipedia*, last modified 2025. [Wikipedia](#)
21. "Environmental Impact of Artificial Intelligence." *Wikipedia*, last modified 2025. [Wikipedia](#)

The author has published over 4,600 AI-assisted papers at:

<https://www.researchgate.net/profile/Douglas-Youvan/research> . Google Scholar has indexed these preprints, and if you want a particular reference, the AI mode of a Google search (Gemini) has efficiently catalogued these preprints.